scientific reports

OPEN

No species-level losses of s2m suggests critical role in replication of SARS-related coronaviruses

Clément Gilbert[1] & Torstein Tengs[2]✉

The genetic element s2m has been acquired through horizontal transfer by many distantly related viruses, including the SARS-related coronaviruses. Here we show that s2m is evolutionarily conserved in these viruses. Though several lineages of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) devoid of the element can be found, these variants seem to have been short lived, indicating that they were less evolutionary fit than their s2m-containing counterparts. On a species-level, however, there do not appear to be any losses and this pattern strongly suggests that the s2m element is essential to virus replication in SARS-CoV-2 and related viruses. Further experiments are needed to elucidate the function of s2m.

The coding capacity of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been investigated in great detail[1], and the secondary structure of genomic RNA elements has also been studied[2,3], but the biological significance of all of these components has not yet been fully elucidated. One of the annotated elements in the reference SARS-CoV-2 genome is the stem-loop II (s2m) element (Genbank accession NC_045512.2, position 29,728–29,768) that was originally described in astroviruses[4]. s2m is a 41-bp sequence located in the non-coding 3′ part of the SARS-CoV-2 genome. It has been found in members of a least four different virus families, including several lineages of coronaviruses[5,6]. There also seems to be a xenolog of s2m in some insect species, which likely results from endogenization of s2m-containing viral elements[7]. The evolutionary relationships between these homologs remain unclear, but it appears as if s2m has been horizontally transferred between distantly related organisms several times[6]. The function is unknown, but the high degree of conservation is consistent with this locus being under selective pressure.

Phylogenetic analyses support several acquisitions of s2m within the coronavirus family, with one gain basal to a cluster of SARS-related betacoronaviruses[5]. This cluster encompasses both SARS-CoV and SARS-CoV-2, as well as many related virus species, primarily isolated from bat species[7,8]. We have done a comprehensive phylogenetic analysis in order to map the distribution of s2m within the *Coronaviridae* subfamily (CoV). In particular, we have tried to assess whether there have been any losses of s2m within the clusters where this motif can be found, with emphasis on the SARS-related species.

All CoV nucleotide and amino acid sequence data were downloaded from GenBank. Based on an alignment of protein sequences from distantly related CoV species, two regions within the ORF1ab polyprotein were identified that could reliably be aligned across a broad range of accessions. The corresponding amino acid sequences from the reference SARS-CoV-2 genome (NC_045512.2 coding positions 10,334–13,468 and 13,462–21,552) were used as query sequences in tblastn sequence similarity searches against the CoV nucleotide data. When tabulating the results, the best matching sequence for every unique GenBank 'ORGANISM' identifier was extracted (Supplementary Table 1). In order to score a viral species as having s2m, the motif had to be found near the 3′ end of the genome with a maximum of one mismatch compared to published s2m sequences[4–7,9] in at least one accession from the corresponding 'ORGANISM' identifier. To remove redundancy in the 436 CoV amino acid sequences that were retrieved from GenBank while retaining their full phylogenetic diversity, we aligned them using MAFTT[10] and removed ambiguously aligned blocks with GBLOCKS[11]. We then used mothur[12] to clusterize s2m-containing sequences and sequences devoid of s2m at 0.1% and 2.5% distance threshold, respectively. We chose to use a higher clustering threshold for sequence devoid of s2m because these sequences were not the focus of our study and were thus primarily included to place s2m-containing sequences in their evolutionary

[1]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, 91198 Gif-sur-Yvette, France. [2]Section of Molecular Toxicology, Department of Environmental Health, Norwegian Institute of Public Health, Oslo, Norway. ✉email: torstein.tengs@fhi.no
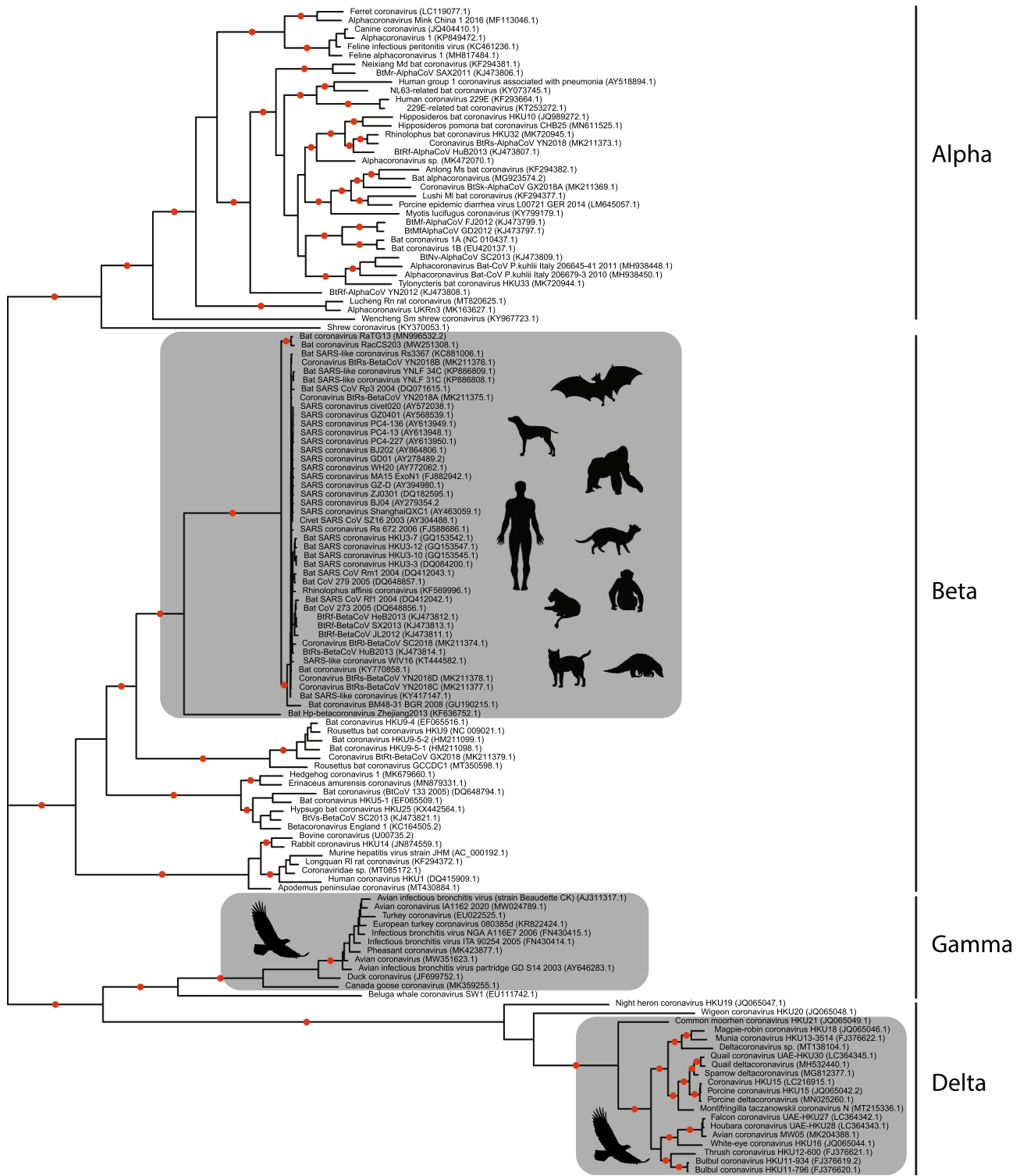
Figure 1. Unrooted maximum-likelihood tree of mothur-clusterized coronavirus ORF1ab sequences. Terminal edges represent single isolates or clusters of highly similar sequences represented by a random sequence within the cluster (see main text and Supplementary Table 1 for details). In addition to the data downloaded from GenBank, the analysis also included sequences from GISAID for strains isolated from lesser known hosts (i.e. cats, dogs, etc., see Supplementary Table 1 for details). Grey boxes represent s2m-containing accessions and *Coronaviridae* genera names are shown. Red dots indicate branches with 100% bootstrap support.
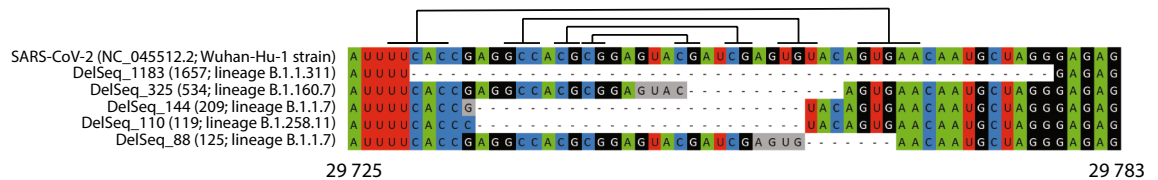
Figure 2. Alignment of the five most common SARS-CoV-2 s2m indel mutants in the GISAID database. Brackets indicate stem-forming regions as described for SARS-CoV[9] and the Wuhan-Hu-1 strain has been included as a reference (nucleotide positions below indicate Wuhan-Hu-1 genome coordinates). Number of instances in the GISAID database and dominant lineage have been indicated (parenthesis after sequence name) and grey boxes show nucleotide(s) positions where there are multiple equally parsimonious ways of making the alignment.

context. The resulting alignment of 133 amino acid sequences was subjected to a phylogenetic analysis using PHYML 3.0[13] with the LG + G + I model, as determined by ProtTest 3[14].

The resulting unrooted topology (Fig. 1) revealed three monophyletic clusters of s2m-containing operational taxonomic units (OTUs). The tree was highly supported, and in addition to two s2m-containing clades comprising isolates stemming from birds, a large group of SARS-related s2m-containing OTUs could readily be identified. This cluster included sequences sampled from several different bat species in addition to eight other vertebrates (Fig. 1). The most basal member of this cluster was Bat Hp-betacoronavirus Zhejiang2013, the only member thus far described from the *Betacoronavirus* subgenus *Hibecovirus*[15]. After excluding a small number of accessions without coverage in the 3′-end of the genome, this clade represented 183 unique ORGANISM identifiers (collapsed into 44 mothur-defined groups, see Supplementary Table 1) that all contained s2m.

Though s2m showed no species-level losses within any of the three clusters, the vast amount of sequence data available from SARS-CoV-2 isolates permit a detailed analysis of how this motif might behave on a virus lineage-level. Sequence data and corresponding metadata from 537 360 SARS-CoV-2 isolates were downloaded from the GISAID database[16]. The 3′ end of high-quality genomes was screened for the presence of s2m single nucleotide polymorphisms (SNPs) and indels. A large number of SNP variants were observed, and, as expected, many of these correlated strongly with virus lineages (as defined by PANGOLIN annotation; Supplementary Table 2)[17]. Looking at indel variants, there also appeared to be lineage-specific variability and several isolates with complete deletion of s2m were observed (Fig. 2; Supplementary Table 3). Two lineages (B.1.1.311 and B.1.160.7) were found to be dominated by s2m deletion mutants (representing 63.7% and 76.1% of the submitted sequences, respectively). Lineage B.1.1.311 had complete deletion of the entire s2m region, whereas lineage B.1.160.7 had a smaller lesion (Fig. 2; DelSeq_1183 and DelSeq_325). Both lineages seem to have had peak distribution Fall 2020 and to have emerged within the United Kingdom (https://cov-lineages.org/lineages/lineage_B.1.1.311.html and https://cov-lineages.org/lineages/lineage_B.1.160.7.html). These lineages have obviously been viable, but their subsequent decline could imply that they were less fit than other emerging strains. Phylogenetic analyses of lineages containing s2m deletion mutants indicated that the primary genetic lesion often is the deletion of a small section of s2m, followed by complete elimination of the element from the lineage's genome (data not shown).

Coronaviruses have been studied extensively in order to identify regions under selective pressure[18]. A recent study identified s2m as having the highest mutation rate in the SARS-CoV-2 genome and the authors suggest that this could be interpreted as either loss of purifying constraints or the result of diversifying selection[19]. Some early reports also proposed that s2m could be involved in recombination events[20]. It is reasonable to assume that the function of s2m is tightly linked with the element's secondary structure. Assuming that the structure is not dependent on interactions with factors that have yet to be identified, an analysis of the canonical SARS-CoV-2 genome using an in vivo-based approach indicated that the structure of s2m deviates significantly from the structure observed for SARS-CoV[3]. The two versions of s2m differ in two positions, constituting two transversions that both seem to disrupt the stem-forming ability of s2m[3]. It is thus unclear if s2m in SARS-CoV and SARS-CoV-2 are functionally equivalent.

In our opinion, the fact that this element never seems to be lost at the species level within the SARS-related coronaviruses suggests that s2m became essential to virus replication after being acquired through horizontal transfer. Both cellular genes and non-coding RNAs acquired by double-stranded DNA viruses through horizontal transfer have been shown to become fixed in viral species, most likely due to their positive effect on viral replication[21–24]. On the contrary, populations of the AcMNPV baculovirus continuously receive transposable elements (TE) from their moth hosts, but all TE copies integrated into the viral genomes become rapidly lost, probably because they impose a fitness cost to the virus[25,26]. We argue that for s2m to be non-essential for viral replication, its distribution within the SARS-related coronaviruses should be significantly more patchy, due to frequent losses.

Though the exact function of s2m has not yet been revealed, several hypotheses have been entertained, including interference with the translational machinery of infected cells[9], involvement in gene regulation through RNA silencing[6] and protection of genomic/transcriptomic virus RNA from ribonuclease degradation[5]. The effect of swapping the 3′ UTR regions between s2m-containg and s2m-deficient virus strains has also been investigated, but the effect on viral fitness seems subtle[27]. Due to s2m's highly conserved nature, it has also been suggested as a potential drug target[28,29] and as a polymerase chain reaction (PCR) primer site for exploring virus diversity[30].

Further studies are needed in order to elucidate the role of s2m in virus replication, not just within the corona-viruses, but in all virus families where this horizontally transferred element has been detected.

References

1. Finkel, Y. *et al.* The coding capacity of SARS-CoV-2. *Nature* **589**, 125–130. https://doi.org/10.1038/s41586-020-2739-1 (2021).
2. Wacker, A. *et al.* Secondary structure determination of conserved SARS-CoV-2 RNA elements by NMR spectroscopy. *Nucleic Acids Res.* **48**, 12415–12435. https://doi.org/10.1093/nar/gkaa1013 (2020).
3. Huston, N. C. *et al.* Comprehensive in vivo secondary structure of the SARS-CoV-2 genome reveals novel regulatory motifs and mechanisms. *Mol. Cell* **81**, 584-598 e585. https://doi.org/10.1016/j.molcel.2020.12.041 (2021).
4. Monceyron, C., Grinde, B. & Jonassen, T. O. Molecular characterisation of the 3′-end of the astrovirus genome. *Arch. Virol.* **142**, 699–706. https://doi.org/10.1007/s007050050112 (1997).
5. Tengs, T. & Jonassen, C. M. Distribution and evolutionary history of the mobile genetic element s2m in coronaviruses. *Diseases* https://doi.org/10.3390/diseases4030027 (2016).
6. Tengs, T., Kristoffersen, A. B., Bachvaroff, T. R. & Jonassen, C. M. A mobile genetic element with unknown function found in distantly related viruses. *Virol. J.* **10**, 132. https://doi.org/10.1186/1743-422x-10-132 (2013).
7. Tengs, T., Delwiche, C. F. & Monceyron Jonassen, C. A genetic element in the SARS-CoV-2 genome is shared with multiple insect species. *J. Gen. Virol.* https://doi.org/10.1099/jgv.0.001551 (2021).
8. Dimonaco, N. J., Salavati, M. & Shih, B. B. Computational analysis of SARS-CoV-2 and SARS-like coronavirus diversity in human, bat and pangolin populations. *Viruses* https://doi.org/10.3390/v13010049 (2020).
9. Robertson, M. P. *et al.* The structure of a rigorously conserved RNA element within the SARS virus genome. *PLoS Biol.* **3**, e5. https://doi.org/10.1371/journal.pbio.0030005 (2005).
10. Katoh, K., Rozewicki, J. & Yamada, K. D. MAFFT online service: Multiple sequence alignment, interactive sequence choice and visualization. *Brief Bioinform.* **20**, 1160–1166 (2019).
11. Talavera, G. & Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577 (2007).
12. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75**, 7537–7541 (2009).
13. Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307–321 (2010).
14. Darriba, D., Taboada, G. L., Doallo, R. & Posada, D. ProtTest 3: Fast selection of best-fit models of protein evolution. *Bioinformatics* **27**, 1164–1165 (2011).
15. Wu, Z. *et al.* ORF8-related genetic evidence for Chinese horseshoe bats as the source of human severe acute respiratory syndrome coronavirus. *J. Infect. Dis.* **213**, 579–583 (2016).
16. Elbe, S. & Buckland-Merrett, G. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Glob. Chall.* **1**, 33–46. https://doi.org/10.1002/gch2.1018 (2017).
17. Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* **5**, 1403–1407. https://doi.org/10.1038/s41564-020-0770-5 (2020).
18. Forni, D., Cagliani, R., Clerici, M. & Sironi, M. Molecular evolution of human coronavirus genomes. *Trends Microbiol.* **25**, 35–48. https://doi.org/10.1016/j.tim.2016.09.001 (2017).
19. Chiara, M., Horner, D. S., Gissi, C. & Pesole, G. Comparative genomics reveals early emergence and biased spatio-temporal distribution of SARS-CoV-2. *Mol. Biol. Evol.* https://doi.org/10.1093/molbev/msab049 (2021).
20. Yeh, T. Y. & Contreras, G. P. Emerging viral mutants in Australia suggest RNA recombination event in the SARS-CoV-2 genome. *Med. J. Aust.* **213**, 44 (2020).
21. Holzerlandt, R., Orengo, C., Kellam, P. & Alba, M. M. Identification of new herpesvirus gene homologs in the human genome. *Genome Res.* **12**, 1739–1748. https://doi.org/10.1101/gr.334302 (2002).
22. Guo, Y. E., Riley, K. J., Iwasaki, A. & Steitz, J. A. Alternative capture of noncoding RNAs or protein-coding genes by herpesviruses to alter host T cell function. *Mol. Cell* **54**, 67–79. https://doi.org/10.1016/j.molcel.2014.03.025 (2014).
23. Theze, J., Takatsuka, J., Nakai, M., Arif, B. & Herniou, E. A. Gene acquisition convergence between entomopoxviruses and baculoviruses. *Viruses-Basel* **7**, 1960–1974. https://doi.org/10.3390/v7041960 (2015).
24. McFadden, G. & Murphy, P. M. Host-related immunomodulators encoded by poxviruses and herpesviruses. *Curr. Opin. Microbiol.* **3**, 371–378. https://doi.org/10.1016/S1369-5274(00)00107-7 (2000).
25. Gilbert, C. *et al.* Continuous influx of genetic material from host to virus populations. *PLoS Genet.* **12**, e1005838. https://doi.org/10.1371/journal.pgen.1005838 (2016).
26. Gilbert, C. & Cordaux, R. Viruses as vectors of horizontal transfer of genetic material in eukaryotes. *Curr. Opin. Virol.* **25**, 16–22. https://doi.org/10.1016/j.coviro.2017.06.005 (2017).
27. Goebel, S. J., Taylor, J. & Masters, P. S. The 3′ cis-acting genomic replication element of the severe acute respiratory syndrome coronavirus can function in the murine coronavirus genome. *J. Virol.* **78**, 7846–7851. https://doi.org/10.1128/Jvi.78.14.7846-7851.2004 (2004).
28. Lulla, V. *et al.* Targeting the conserved stem loop 2 motif in the SARS-CoV-2 genome. *J. Virol.* **95**, e0066321. https://doi.org/10.1128/JVI.00663-21 (2021).
29. Manfredonia, I. *et al.* Genome-wide mapping of SARS-CoV-2 RNA structures identifies therapeutically-relevant elements. *Nucleic Acids Res.* **48**, 12436–12452. https://doi.org/10.1093/nar/gkaa1053 (2020).
30. Jonassen, C. M. Detection and sequence characterization of the 3′-end of coronavirus genomes harboring the highly conserved RNA motif s2m. *Methods Mol. Biol.* **454**, 27–34. https://doi.org/10.1007/978-1-59745-181-9_3 (2008).

Author contributions

The project was conceptualized by T.T. C.G. did the all the phylogenetic analyses and the final version of the manuscript was written by both authors.

Funding

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-95496-4.

Correspondence and requests for materials should be addressed to T.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.